EVALUATION DES TESTS DE DIAGNOSTIC

Plan de la présentation

- Introduction
  - Définition Dépistage, Diagnostique
- Les ¾ phases de développement d'un test
  - Phases 1, 2, 3 et 4
- Les indices de performances
  - Reproductibilité (Indice Kappa de Cohen, diagramme de Bland et Altman, CCI)
  - Validité (Se, Sp, VPP, VPN, LR+, LR-)
  - Courbes Roc
- Les principaux biais

Les critères diagnostiques

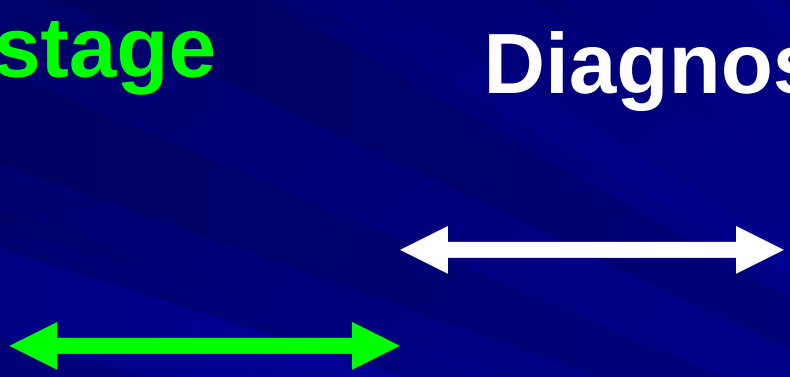
- Ce sont des marqueurs dont les résultats permettent d'orienter la décision médicale
- Deux niveaux:
  - Les tests de dépistage
  - Les tests de diagnostic
- Comprennent

  Test médical, signes cliniques, critères bactériens, images radiologiques, tests biochimiques ..

- Un test peut être **très simple** par exemple **une culture bactérienne** pour évaluer une infection

- Ou **plus compliqué** par exemple un **score clinique** à partir d'un questionnaire ou **la séquence de procédures spécifiques** selon un protocole

Dépistage et diagnostic



exposition	Phase pré-clinique	Phase clinique	Complications

Diagnostic et Dépistage

Le diagnostic se distingue du **dépistage** par une caractéristique fondamentale :

La motivation de l'examen

- Réalisé en raison de l'état clinique (sujet malade)

  ➔ **diagnostic**

- Réalisé indépendamment de l'état clinique (sujet apparemment sain)

  ➔ **dépistage**

L'évaluation statistique est identique que ce soit pour

un Test de « dépistage »

ou

un Test de « diagnostic »

Les 3/4 phases de développement d'un test diagnostique

La phase I (proof-of-concept)

L'objectif de cette phase exploratoire est de vérifier que les résultats du test sont différents chez les malades et les non malades (conditions « de laboratoire »)

Vérification du « mécanisme » d'action

Test dans conditions différentes (températures, humidité..), conditions de recueil

Test chez sujets ou échantillons différents (différents niveaux de sévérité de maladie ou de quantité de bactéries….) y compris des sujets non malades mais avec des symptômes proches de ceux des sujets malades

Test de la reproductibilité

Exemple: Bandelettes pour diagnostiquer des shigelles à partir de selles

- **Vérifier**
  - **bandelettes + dans prélèvements contenant des shigelles**
  - **Bandelettes - dans prélèvements sans shigelles**
- **Définir les conditions de recueil des selles (délai de recueil et conditions de stérilité..)**
- **Vérifier délai et conditions de lecture (mode d'emploi)**
- **Vérifier la reproductibilité de la lecture (utiliser 2 bandelettes, lues par deux lecteurs différents)**
- **Faire varier les conditions de stockage des bandelettes (humidité, température)**

➔ **Cette phase permet de savoir si le test semble suffisamment « fiable » pour pouvoir être utilisé dans les conditions souhaitées**

Une bonne reproductibilité est nécessaire

La phase II (Etude cas-témoins)

L'objectif de cette phase de validation est de montrer que

- **la probabilité d'avoir un résultat + est supérieure chez les malades**
- **la probabilité d'avoir un résultat – est supérieure chez les non malades**

➔ **Validité du test dans conditions contrôlées (≠ conditions de terrain)**

- **Sélectionner les cas et les témoins, les évaluateurs (médecins, infirmiers, radiologues..) qui peuvent différer du terrain**

- **Définir dans un PROTOCOLE les conditions d'utilisation du test (en essayant d'éviter les biais)**

- **Tester différentes conditions de recueil (températures, humidité..)**

- **Estimer le % de faux positifs et de faux négatifs (calcul du NSN et estimations des valeurs acceptables)**

- **Pour les tests quantitatifs, définir le cut-off (Courbes Roc), identifier les facteurs ayant un impact sur le test (ou ceux qui le rendent ininterprétable)**

- **Les études de phases I et II sont des études rétrospectives réalisées uniquement dans un but de recherche**

- **Le statut du malade est déterminé avant par d'autres moyens**

La phase III (Etude prospective)

Objectif principal: Déterminer les performances du test dans les conditions où il sera utilisé

➔ vérifier que chez les patients chez lesquels il est cliniquement pertinent, les résultats du test permettent de distinguer les malades des non malades

Les performances pourront être comparées à celles d'autres tests

➔ Réalisée dans les conditions pratiques d'utilisation du test

➔ S'adresse à des sujets dont on ne connaît pas à l'avance l'état (Malade ou Non Malade)

Phase IV ?

Idéalement, il est intéressant de comparer, par un essai randomisé, l'impact de l'introduction du test par rapport à une prise en charge sans test, dans la pratique courante, sur des critères de morbi-mortalité (qualité de vie) et de coûts

➔ Permet de savoir si les sujets « testés » se « portent mieux » que les sujets « non testés »

Relation entre le Taux de peptide natriurétique (PN) et hypertrophie ventriculaire gauche (HVG)

Phase I

Les patients avec une HVG ont-ils des concentrations de PN supérieures à celles observées chez des sujets normaux?

	Patients HVG +	Patients HVG -
PN (pg/ml) (moyenne)	493.5	129.4
médiane (range)	(248.9-909.0)	(53.6-159.7)

Relation entre le Taux de peptide natriurétique (PN) et hypertrophie ventriculaire gauche (HVG)

Phase II

Les patients avec des concentrations de PN élevées ont ils plus souvent une HVG que ceux avec des concentrations faibles ?

PN (pg/ml)	Cas (Avec HVG)	Témoins (Sans HVG)
Valeurs élevées	n=39	n=2
Valeurs normales	n=1	n=25

Se= 98% (87-100) **VPP=95% (84-99)**

Sp= 92% (77-98) **VPN=95% (84-99)**

LR+= 13 (3.5-50.0) **LR-= 0.03 (0.0003-0.19)**

Relation entre le Taux de peptide natriurétique (PN) et hypertrophie ventriculaire gauche (HVG)

Phase III

Parmi les sujets chez qui une suspicion clinique de HVG existe, les taux de PN sont-ils différents entre ceux qui ont une HVG (Echo) et ceux qui n'en n'ont pas ?

PN (pg/ml)	Sujets HVG +	Sujets HVG -
Valeurs élevées (>=18)	n=35	n=57
Valeurs normales (<18)	n=5	n=29

Se = 88% (74-94) **VPP**=38% (29-48)
Sp = 34% (25-44) **VPN**=85% (70-94)
LR+= 1.3 (1.1-1.6) **LR-** = 0.4 (0.2-0.9)

Relation entre le Taux de peptide natriurétique (PN) et hypertrophie ventriculaire gauche (HVG)

Phase IV

Les sujets chez qui une suspicion clinique de HVG existe et chez qui un dosage de PN a été réalisé ont - ils « un meilleur état de santé » que les sujets qui n'ont pas été testés ?

Indices de performances

Reproductibilité

Validité (Accuracy)

Validité d'un test par rapport à un Gold Standard

Définition du Gold Standard

- Caractéristiques des critères permettant d'affirmer l'existence d'une maladie
- Un critère indiscutable: tuberculose et BK ?
- Tuberculose et IDR+?
- Critères histo-pathologiques à partir de Biopsies

Difficultés du Gold Standard

- **Pas de gold standard : aucun critère vraiment satisfaisant**

  Ex= test rapide de bandelettes pour diagnostiquer shigelles chez enfants atteints de diarrhée sévère (Coproculture peu sensible, PCR non spécifique)

- **Quand un nouveau test ferait mieux que le standard actuel**

  Ex: bandelettes plus sensibles si lues rapidement

- **Un gold standard ne doit comporter dans sa définition, ni le signe, ni le résultat du test dont on évalue les propriétés diagnostiques**

Le test

- **Les critères de positivité d'un test doivent être connus précisément**

➔ **Décrire les conditions dans lesquelles ils sont mesurés et les règles de conclusion**

Exemples

- Utilisation d'une bandelette pour diagnostiquer des shigelles dans les selles ➔ lecture dans les 15 minutes …
- Mesure de la glycémie: à jeun, post prandiale
- Examen direct BK (conditions de lecture : nombre de champs microscopiques?)

Evaluation d'un nouveau test diagnostic et critères de performances

- **On peut distinguer**

  - **Les tests binaires (oui/non ou positif/négatif ou présent/absent)**

  Ex: Présence de sang dans les urines, sérologie VIH positive ou négative, BAAR + ou –

  - **Les tests quantitatifs: variable continue avec un seuil**

  Ex: bilirubinémie, cholestérolémie, taux de PSA, FibroScan

  - **Les réponses ordinales**

  Ex: degré de fibrose sur une lame de biopsie, images radiologiques, échelle de BIRADS sur la mammographie

  ➔ analyse peut être abordée comme du quantitatif

En fonction du critère, la méthodologie d'évaluation sera différente

Expressions des résultats d'une évaluation

- Signe binaire : Sensibilité, spécificité
- Signe avec valeur continue : Courbes ROC

Cas des variables binaires

Expression des résultats

Statut / Résultat	malade M+	non malade M-
test +	Vrai Positif VP	Faux positif FP
test -	Faux Négatif FN	Vrai Négatif VN

On distingue 4 types de sujets
-Les vrais positifs (VP)
-Les faux positifs (FP)
-Les varis négatifs (VN)
-Les faux négatifs (FN)

Qualités intrinsèques : sensibilité et spécificité

	M+	M-
T +	VP	FP
T -	FN	VN

Se = P (T+ / M+) = VP / VP+FN

Sensibilité : probabilité d'obtenir un test positif quand le sujet est malade

Valeur comprise entre 0 et 1

=> c'est l'aptitude d'un test à identifier correctement les individus malades grâce à une réponse positive

Qualités intrinsèques : sensibilité et spécificité

	M+	M-
T +	VP	FP
T -	FN	VN

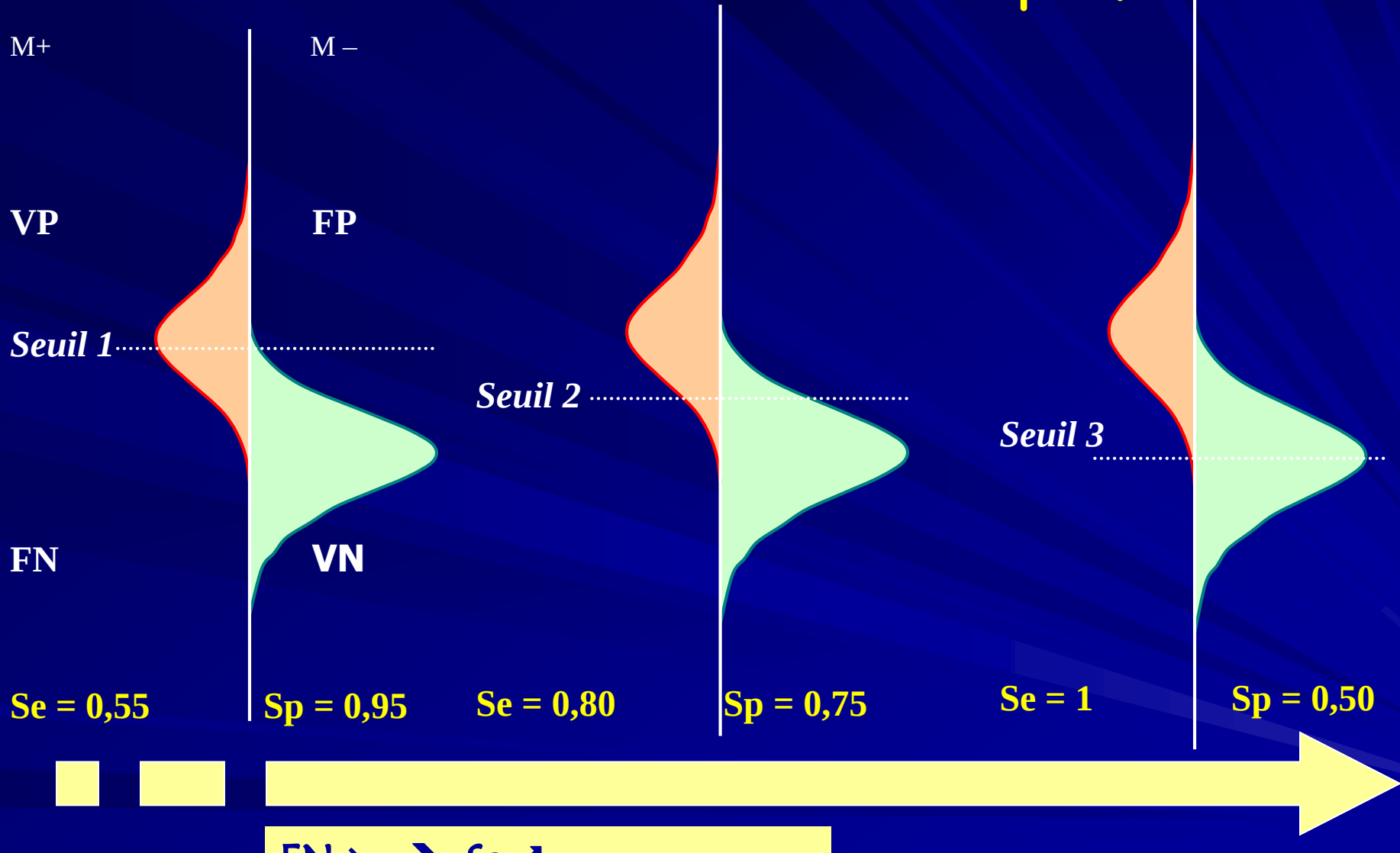
$Sp = P(T- / M-) = VN / VN+FP$

Spécificité : probabilité d'obtenir un test négatif quand le sujet est non malade

Valeur comprise entre 0 et 1

=> c'est l'aptitude d'un test à identifier correctement les individus non malades grâce à une réponse négative

Relativité de la sensibilité et de la spécificité



FN ↘ ➔ Se ↗

FP ↗ ➔ Sp ↘

Ex: dépistage cancer du sein

- HIP Breast Cancer Screening Project
- 64810 femmes âgées de 40 à 64 ans

Ex. physique + mammo.

	Cancer du sein (biopsie ou aspiration) +	-	
+	**132**	**983**	**1115**
-	**45**	**63650**	**63695**
	177	**64633**	**64810**

Sensibilité: 132/177 = 75%

Spécificité: 63650/64633 = 99%

(Shapiro S et al., Am J Epidemiol, 1974)

Cas d'un signe avec des valeurs quantitatives

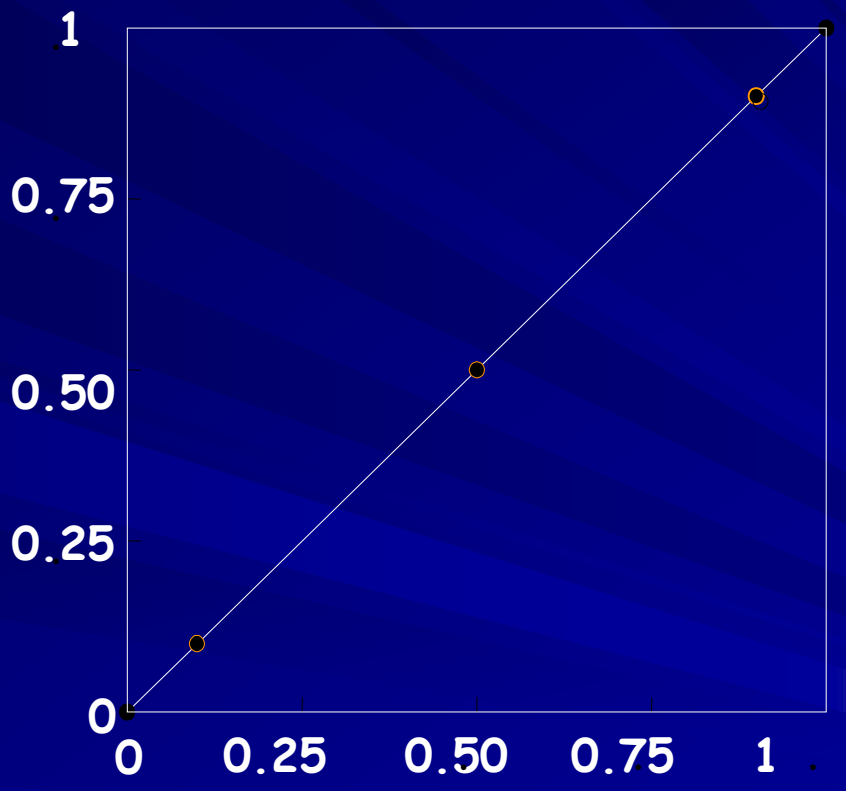
Les courbes ROC

Lorsqu'un test a des valeurs continues, il existe de nombreuses valeurs Se/Sp

➔ important d'avoir tous les spectres des valeurs du test et de ne pas se limiter à certaines valeurs, ou intervalles même si le but final est de définir un seuil (cut-off) qui présente le meilleur rapport Se/Sp (celui qui nous intéresse et qui dépend du contexte)

La courbe ROC permet d'avoir le tracé des Se et Sp correspondant à toutes les valeurs du test

Sensibilité (Vrais positifs)



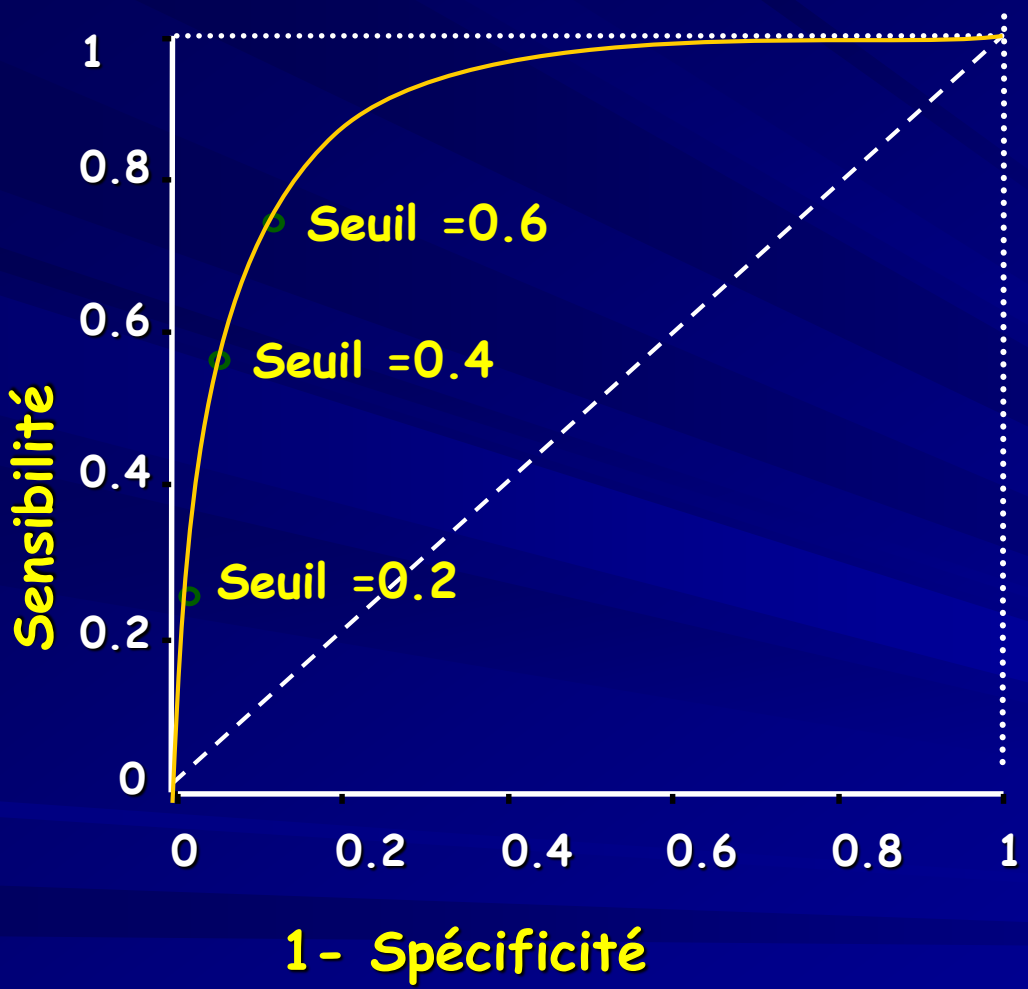
1-Spécificité (Faux positifs)

Construction de la courbe ROC

On porte

- en abscisse, la variable « 1 – spécificité » = faux positifs parmi les non-malades
- en ordonnée la sensibilité = les vrais positifs parmi les malades

Un seuil est idéal s'il permet de séparer totalement les positifs des négatifs, sans faux positifs ni faux négatifs



$p_1 > S$

$p_1 \leq S$

M+	M-
VP	FP
FN	VN

Seuil 0.2

$p_1 > S$

$p_1 \leq S$

M+	M-
VP	FP
FN	VN

Seuil 0.4

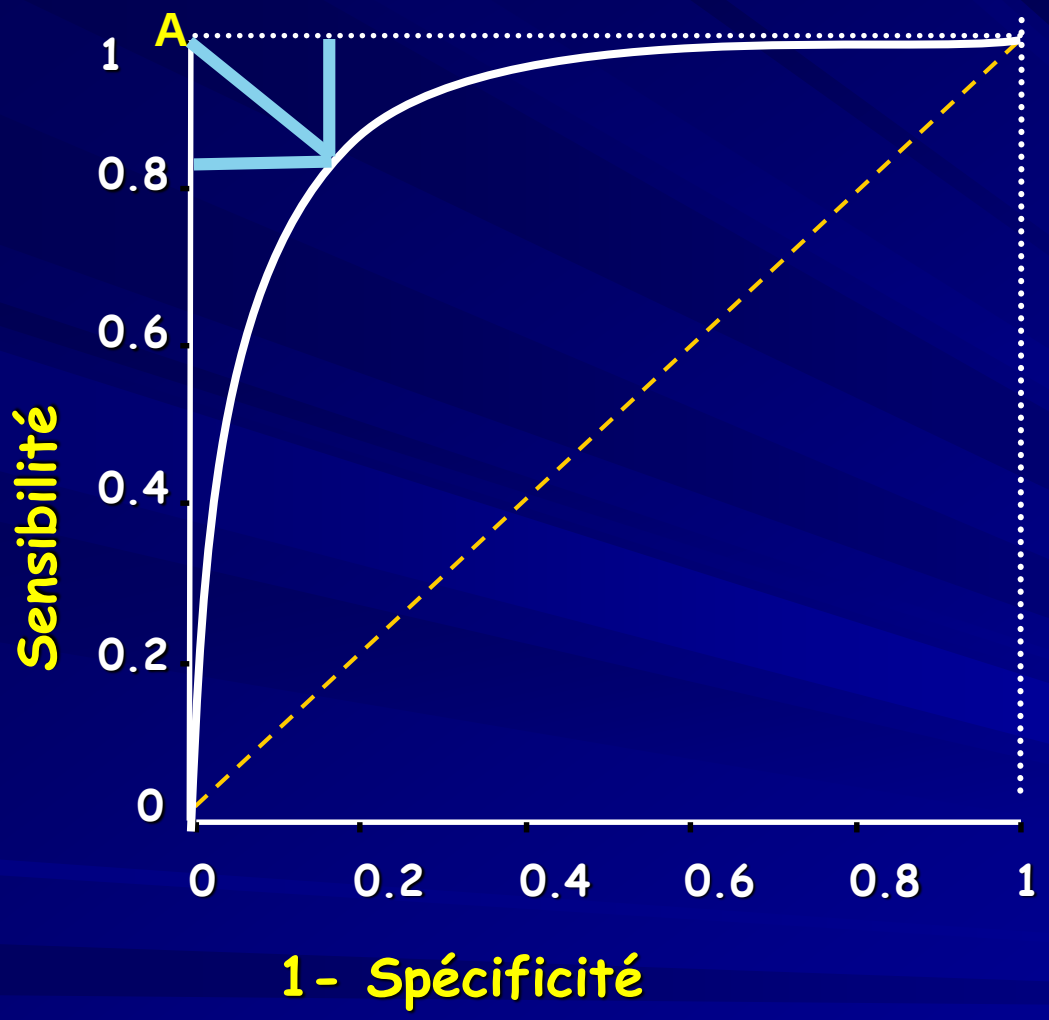
$p_1 > S$

$p_1 \leq S$

M+	M-
VP	FP
FN	VN

Seuil 0.6

Construction de la courbe ROC



Diagonale passant par 0
➔ test non informatif

➔Un test est d'autant meilleur que sa courbe se situera près du point A et loin de la diagonale

➔Minimiser la distance par rapport au point idéal A

➔ Se = Sp = 1

- On calcule AUC= Aire sous courbe (avec $IC_{95\%}$)

- Plus AUC proche de 1 plus le test est performant

Avantages des COURBES ROC

- Simple et facilement compréhensible graphiquement
- Tient compte de l'ensemble des valeurs du test (ne nécessite pas le choix arbitraire d'un seuil)
- Totalement indépendante de la prévalence de la maladie dans l'échantillon
- Permet une comparaison directe visuelle de plusieurs tests sur une même échelle (+ tests pour comparer AUC)
- on peut calculer l'$IC_{95\%}$ de l'AUC. La borne inférieure ne doit pas comprendre 0.5 pour que le test ait un intérêt

Les principaux biais

Definitions

- **Erreur systematique (Biais) : erreur de raisonnement ou de procédure amenant à une représentation faussée de la réalité. Un biais revêt un caractère systématique et altère l'estimation dans un sens donné.**

- **Erreur aleatoire : non imputable à une cause décelable et due au hasard. Elle aboutit à une perte de précision de l'estimation mais non à sa déviation systématique dans un sens donné.**

6) Sources de BIAIS….

A) Intégrité des tests (Integrity test)
B) Le biais de vérification (Verification bias)
C) Erreurs sur la référence (Errors in the reference)
D) Le biais de recrutement (Spectrum bias)
E) Le biais d'interprétation (Test interpretation bias)
F) Les tests ininterprétables (Unsatisfactory tests)
G) Le biais d'extrapolation (Extrapolation bias)
H) Le biais d'incorporation (Incorporation bias)

+ biais spécifiques des études de dépistage
I) Le biais du temps d'avance au diagnostic (Lead Time Bias)
J) Le biais de lenteur d'évolution (Length Time Bias)
K) Le biais de sur-diagnostic (Diagnostic Bias)
L) Le bais de sélection (Selection Bias)

A) Qualité des tests (Integrity test)

Il est nécessaire que:

- **La connaissance du statut de la maladie (OUI/NON) des sujets n'influence pas l'évaluation du test (et vice versa)**

  Par exemple: si un radiologue sait que la mammographie qu'il doit évaluer vient d'une femme atteinte d'un K du sein, il pourra être influencé

  ➔ les « opérateurs » qui évaluent le test doivent être à l'insu du résultat du statut de la maladie (vice versa)

- **Souvent les procédures d'évaluation «objectives » et l'évaluation de la maladie n'interfèrent pas (questionnaire ou évaluation par un médecin # test biochimique ou test sur culture)**

  ➔ on dit que « the integrity of such test is inherent to its operation ».

- **Mais même dans le cas de tests « objectifs », il faut se méfier**

- **Exemple 1**

  Si le délai entre le test et l'évaluation de la maladie est long, il peut y avoir une modification de la prise en charge qui est fonction du résultat du test et ainsi entraîner une modification du diagnostic de la maladie ➔ les résultats seront biaisés

- **Exemple 2**

  A l'inverse, la connaissance de la maladie peut influencer la façon dont le test sera réalisé ➔ les résultats seront biaisés

B) Le biais de vérification (Verification Bias,Work-up Bias,Referral bias, selection Bias ou Ascertainment Biais)

Dans les études de cohortes, le test T devrait être appliqué à tous les sujets

Mais…on peut être dans la situation où la référence n'est réalisée que si le test est +

- ➔ Si T+ ➔ On réalise un test de référence pour confirmer la maladie
- ➔ Si T - ➔ Pas de test de référence

Exemple

Détection test audition des bébés à la naissance par DPOAE (Test d'émission oto-acoustique : valeur normale chez l'adulte = 100)

- **Si le résultat n'est pas parfait** on fait le test de référence (VRA (Visual Reinforcement Audiometry)
- **Si le résultat est bon** on ne fait pas le test VRA qui est cher, long et nécessite un second RV

C) Erreurs sur la référence (Imperfect Reference Test)

- Pour de nombreuses maladies, il est impossible de déterminer avec certitude le statut de la maladie et les meilleures références peuvent donner des résultats faux (PBH)

➔ Ces erreurs peuvent avoir des conséquences sur l'évaluation d'un test diagnostic

Ex: **Infection bactérienne**: culture à partir d'un prélèvement de sang, urine ou tissu, même si le sujet est infecté, la culture peut être – si le spécimen ne contient pas le germe ou si l'échantillon contient la bactérie mais ne pousse pas

- **La PBH** (résultat dépend de l'endroit où le prélèvement est effectué, la taille du prélèvement etc..)

Il existe des méthodes d'analyses (**analyse avec classes latentes**) pour Évaluer différents tests en cas d'absence de Gold standard

D) Le biais de recrutement (Spectrum Bias)

- Quand les sujets malades ne sont pas représentatifs des sujets malades de la population ou quand les sujets contrôles (non malades) ne sont pas représentatifs des sujets non malades de la population

- Erreur classique: sélectionner des cas « graves » et des contrôles « très sains »

➔ paramètres de validité des tests surestimés

E) Le biais d'interprétation

- **Quand des informations «extérieures» (données cliniques ou résultats d'autres tests) peuvent influencer la procédure du test à évaluer qui ne sera pas appliquée ainsi dans la pratique ultérieure**
- **<u>Exemple</u>**

  Les résultats d'une mammographie peuvent influencer l'interprétation d'une «grosseur» lors d'un examen clinique. Si en pratique, l'examen clinique est réalisé sans mammographie, les performances du test peuvent être différentes

F) Les biais liés aux tests ininterprétables

- En pratique les tests ne sont pas applicables à tous les sujets et pour certains sujets ils peuvent être ininterprétables
- **Exemples :**
  - FibroScan chez sujets obèses
  - Test d'audition chez un enfant agité
- Ces informations doivent être prises en compte dans l'évaluation des tests
- Si non prises en compte ➔ surestimation de la validité du test alors que chez certains sujets la maladie n'est pas détectée
- **A l'inverse,** la prise en compte de ces données peut être problématique, par exemple si les «cas non interprétables» sont considérés comme des négatifs avec en pratique des recommandations faites pour répéter le test

  Or les performances du test ne sont pas évaluées pour des valeurs répétées

G) Le biais d'extrapolation

- **Plusieurs facteurs peuvent influencer les performances d'un test pour détecter une maladie**
  - **Facteurs liés à l'opérateur : expérience**
  - **Facteurs liés au sujet: Caractéristiques démographiques**
  - **Environnement dans lequel le test est réalisé : ressources disponibles, accès aux traitements, prévalence de la maladie**

➔ **Les performances d'un test réalisé dans une population ne peuvent pas être systématiquement extrapolées à d'autres populations**

H) Le biais d'incorporation

Lorsque le résultat du test est incorporé dans le diagnostic
Par exemple si on veut savoir si la radiographie est un bon marqueur de TB chez les enfants infectés par le VIH

Le test est la radiographie et la référence est le diagnostic de TB défini par un groupe d'experts qui va statuer sur les signes cliniques, la culture du crachat et la radiographie

En résumé le test étudié doit

- être fait par ceux qui le feront en routine
- être fait à tous les patients étudiés
- ne pas être incorporé dans le gold standard
- être complètement décrit
- être interprété à l'aveugle / gold standard, avec ou sans informations cliniques
- le taux de résultats ininterprétables doit être fourni

Interprétation du test

Les résultats intermédiaires ou indéterminés constituent un **résultat**

Pour les résultats ininterprétables, il faut préciser **les conditions** et essayer de savoir s'il y a une **relation avec le diagnostic**

Conclusion

Pour qu'un critère devienne un test de dépistage ou un test diagnostique

- **Nécessité d'évaluer la reproductibilité, les qualités intrinsèques, extrinsèques…**
- **à partir de protocoles réalisés sans biais**
- **en respectant les 3-4 phases d'évaluation**

- **Références** : Grille **STARD & QUADAS**

- **Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative.** Ann Intern Med 138: 40–44.
- **Whiting P, Rutjes A, Reitsma J, Bossuyt P, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. B**MC Medical Research Methodology 2003;3:25.

Impact des défauts de méthodologie (Reid et al. JAMA, 1995)

- **112 études de 1978- 1993 : 7 standards méthodologiques**

-Définition de la population (âge, sexe, symptômes, critères d'éligibilité) : 27%
-Qualité DG dans différents sous-groupes: 8%
-Absence de biais de vérification : 46%
-Absence de biais d'évaluation (test ou référence) : 38%
-Précisions des estimations : 11%
-Présentations des résultats indéterminés : 23%
-Reproductibilité du test :23%

Performance de plusieurs tests rapides pour le dépistage de la Dengue

Test	Performances déclarées		Evaluation OMS	
	Se	Sp	Se	Sp
Core	100	100	23	99
Diazyme	NS	NS	18	98
Globalemed	80	>99	63	69
Minerva	NS	NS	9	100
Panbio	70	100	65	98
Standard	93	100	22	99
Tulip	100	100	6	99

World Health Organization
An ideal diagnostic test : ASSURED

- A = **Affordable by those at risk of infection**
- S = **Sensitive**
- S = **Specific**
- U = **User-friendly (simple to perform, minimal training)**
- R = **Rapid/robust (enables action at point of care)**
- E = **Equipment-free**
- D = **Deliverable to those who need it**